

The Viability of Using Rapid Judgments as a Method of Deception Detection

Norah E. Dunbar¹, Lindsey A. Harvell², Matthew L. Jensen¹,
Judee K. Burgoon³, Katherine M. Kelley⁴

¹University of Oklahoma, ²Southern Utah University,
³University of Arizona, ⁴Fort Hays State University

Abstract

Rapid Judgments (RJs) are assessments based on indirect verbal and nonverbal cues made on the basis of global judgments. Advantages of RJs, compared to other time-consuming methods of deception detection, include the ability to make assessments quickly, not having to use expensive detection equipment, and only minimal training for coders. Vrij and colleagues [1] found through initial testing with novice observers that the viability of the RJs method in deception detection is promising. The purpose of this research was to test the RJs method with two additional datasets in order to verify the reliability of the coding method and the accuracy of the deception judgments made with minimally trained coders. Two hypotheses were proposed and only partially supported.

1. Introduction

Humans are notoriously poor at detecting deception. A recent meta-analysis found that people are 54% accurate, on average, when deciding between truth and deception which is only slightly above what would be found by chance alone [2]. Levine [3] calls the consistency of deception detection accuracy rates across studies one of the most reliable, consistent, and robust findings in all of social science. Even professional lie detectors such as police officers, unaided polygraph examiners, and judges perform around 56-65% and rarely exceed 70% accuracy [4, 5]. The reasons for poor detection are many (see [5] for a review) but one reason is the cognitive complexity of the detection task. Lie detectors must monitor a host of vocal and nonverbal cues as well as examine the verbal content of a statement for logical inconsistencies. This task is even more difficult when the lie detector is a party to the conversation rather than merely an observer because of the cognitive load associated with maintaining a conversation [6, 7] and more difficult when the deceiver is particularly skilled. Thus, “unaided” deception detection, even when extensive training is involved, has often proven to be an ineffective method of deception detection [8].

2. Literature review

To encourage more accurate deception detection, researchers have created a variety of complex data analysis or coding methods, either through manual human coding

or computer-based assessment, ranging from close analysis of nonverbal cues [9-11], linguistic and verbal analysis of transcripts [12, 13], vocalic analysis of auditory cues [14-17], and analysis of physiological information such as cardiovascular, respiratory, and pupillary responses [18, 19]. These methods have been demonstrated to improve accuracy to over 80% but they are time-consuming and often require laborious hours of training, transcription of statements, or post-interaction analysis, often with the aid of specialized equipment [1, 11]. For example, the polygraph is designed to detect deception through monitoring physiological indicators of stress and arousal, and automated linguistic analysis can record, transcribe, parse, and analyze a conversation in nearly real time. Both of these methods, although their accuracy rates range from 60-90%, require many hours of training and preparation even though they provide real-time data [19, 20].

Training people to detect deception, even when they are professionals, can often be a lengthy and difficult task [21-23]. Accurately coding behaviors, either verbal or nonverbal, often requires an individual to watch a video clip or listen to an audio recording several times until all of the verbal and nonverbal behaviors have been coded [1]. In fact, some behaviors, such as voice pitch, often require special equipment for detection [24]. Researchers have devoted considerable attention to the training that is required for coders or judges of these behaviors in deception detection research [25]. Observers who are trained will often focus on those cues mentioned in the training program while dismissing unmentioned cues [26]. Therefore, training programs tend to bias the observers’ decisions while increasing their judgment confidence [26]. Levine et al. [27] found that even “bogus” training can increase accuracy because, as they argue, the main result of training is an increased vigilance but not necessarily greater accuracy. Masip and colleagues [26] suggest the idea that training should be re-examined in order not to bias observers’ judgments or their judgment accuracy. This can be done by focusing training sessions on deception and truthfulness, enabling the observer to be more open minded and perceive every cue with more accuracy [26]. However, even successful training programs only increase accuracy by 5-10% [25, 28-30]. Thus, if accuracy can be increased with minimally trained coders, such a coding method would be a significant advancement in the study of deception detection.

An alternative to time-intensive coding or computer-aided detection is credibility assessment based on “indirect cues” indicating deception [1, 31-33]. The indirect cues approach, also known as rapid judgments (RJ), asks coders to make global assessments on such aspects as the amount of detail provided by a speaker, rather than make a direct truth-deception judgment. Vrij and colleagues [1] argue that this method of assessment addresses several problems created by the unaided detection method, the manual human coding method, and the computer-based assessment methods described above. First, there is a general reluctance, especially on the part of unaided detectors, to characterize messages as deceptive. A meta-analysis of deception detection studies noted a distinct truth bias in deception judgments [34]. This tendency may be evidence of an availability heuristic or may stem from individuals wishing to extend the benefit of doubt to others who may actually be deceptive [35]. Credibility assessment based on indirect cues mitigates this issue by querying interaction observers about cues that are indicative of deception (e.g., Does this person believe what he is saying?) rather than a direct measure of deception (e.g., Is this person lying?). This avoids the undesirable task of labeling someone a deceiver.

Second, there are widespread misconceptions about what verbal and nonverbal behaviors are correlated with deception. For example, there is widespread belief that eye gaze is indicative of deception [36] but reviews of the deception literature suggest that there is no connection between gaze aversion and deception [9, 35]. Even when nonverbal behaviors can be identified that are related to deception [9, 12, 37], these are not consistently used during credibility assessment by detectors [9, 35]. By assessing credibility via indirect cues (e.g., responding to: “Does this person have to think hard when answering?”), detectors of deception report their observations about behaviors that have been actually linked to deception rather than report observations that are stereotypically (and often incorrectly) associated with deception.

A final advantage of using indirect cues is that assessments based on indirect cues rely on global judgments. These judgments can be made quickly without transcripts or specialized equipment and with minimal training. Given this, rapid judgments are preferable to both the subjective judgments of human lie detectors and the time-consuming coding methods, but only if detection accuracy can be consistently demonstrated. Initial studies testing the viability of RJs found promising results with novice observers [1] and police officers [33].

Our purpose in this research is to test the RJs method with two additional datasets to verify the reliability of the coding method and the accuracy of the deception judgments made with minimally trained coders.

3. Theory and Hypotheses

The RJs approach to coding suggests that when people are asked to pay attention to verbal and nonverbal cues, rather than detect deceit, they are more accurate in their judgments [38]. It is further argued that people should be able to detect deceit immediately following their RJs because they have been focused on diagnostic cues rather than heuristic judgments [1]. After training, research has shown that people do become better lie detectors when they are instructed to look at diagnostic cues [38, 39]. Therefore, by focusing on cues, instead of lies, this method is more successful in detecting deceit [32, 33].

The RJs explored in this work are based on Vrij et al.’s [1] study. The RJ cues used in this study come from a large body of literature on two techniques for the analysis of written transcripts known as Criteria-based Content Analysis (CBCA) and Reality Monitoring (RM). Using CBCA, trained evaluators judge the presence or absence of nineteen criteria that are broken up into four main areas: (1) general characteristics, (2) specific characteristics, (3) motivation-related contents, and (4) offense-specific elements. The first area, general characteristics, consists of the logical structure of a statement, unstructured production of content, and the quantity of details provided. The second area, specific characteristics, includes contextual embedding of detail, descriptions of interactions, reproduction of conversation from memory, unexpected complications during the interview, unusual details that might come up, adding superfluous details, providing accurately reported details that are later misunderstood, related external associations (e.g., “I bragged to my friends about my receiving credit for someone else’s work”), accounts of subjective mental states, and attributions of the hypothetical perpetrator’s mental state. The third area, motivation-related content, focuses on spontaneous corrections, admitting lack of memory, raising doubts about one’s own testimony, self-deprecation, and pardoning the perpetrator. Lastly, offense-specific elements look at details of the characteristics of the defense. Similarly, in RM analysis, evaluators use eight criteria, some of which overlap with CBCA, which are designed to discern between real and false memories including the amount of affect displayed and the amount of temporal and spatial information given (the CBCA and RM criteria are reviewed in greater detail by [35]). Both of these methods require transcription of verbal statements and laborious manual coding by specially trained human coders. The RJ approach draws on CBCA and RM but does not require transcripts or detailed coding.

Vrij et al. [1] examined whether observers a) make rapid, reliable, and valid judgments of the frequency of verbal and nonverbal behaviors of interviewers and b) are able to detect deceit after making RJs. They had five coders view 52 clips of 26 liars and 26 truth tellers using 12

nonverbal cues and some of the CBCA, and RM criteria and then make a truth-deception judgment following their coding. They trained the coders on the RJ coding scheme in approximately 90 minutes of training [1]. We followed their methodology and trained our coders on the 10 RJs found in Table 1, but initial analyses revealed that not all of the RJs were used by the interviewees in our videos, so only the bolded RJs were included in our analyses.

Table 1: Codes and definitions

Code	Definition
Quantity of Details	Specific descriptions of place, time, persons, objects, and events.
Contextual Embeddings	Descriptions of time and location.
Reproduction of Conversation	Speech reported in its original form.
Description of other’s Mental State	Description of other people’s feelings, thoughts, or motives.
Visual Details	Description of details which the interviewee saw.
Auditory Details	Description of details which the interviewee heard.
Spatial Information	Information about locations and about how objects were related to each other.
Temporal Details	Information about time and duration of events
Cognitive Operations	Descriptions of inferences made by the participant at the time of the event.
Spontaneous Corrections	Corrections are made or information is added without prompting by the interviewer.

Following the method of Vrij et al., we analyzed two existing datasets to further test the viability of the RJs approach. The first is a dataset collected by Dr. Burgoon’s research team at the University of Arizona which we will call the “Study 1 – Spontaneous Deception” [40]. In this study, international students from a variety of cultural backgrounds answered questions during an interview conducted by a professional polygraph examiner and were prompted to lie or tell the truth on the question by a teleprompter. The second study, which we will call the “Study 2 – Cheating Denial”, was conducted by Drs. Dunbar and Jensen’s research team at the University of Oklahoma in which students were induced to cheat on a task with a confederate and then were interviewed subsequent to the task by a professional polygraph examiner [41]. Although Vrij et al. [1] had the coders make a dichotomous truth-deception judgment following the

RJs, we had our coders rate the truthfulness of the statement on a 7-point scale in order to be consistent with the interviewers’ and participants’ ratings at the time of the original data collection (see Vrij et al., [1], for review).

The Vrij et al. [1] study demonstrates convincingly that RJs can be used reliably for short clips (interviews 2-3 minutes in length about a “Connect Four” game) but the study does not speak to longer interviews like those used by law enforcement. Our goal was not only to replicate the Vrij et al. [1] study but also to determine if any limits exist for the application of RJs.

There could be several limitations that could occur with longer interviews. First, observers may have a harder time keeping track of every code that occurs in the interview when the segment is long and there is a lot of information being exchanged. Second, there is more of an opportunity for high and low behaviors to cancel each other out. For example, if one part of the interview had a great deal of auditory detail but another part does not, judges may be inclined to choose the middle of the scale. As a result, the longer an interview is, the more possible it becomes for RJs to regress to the middle of the scale. The stimulus materials we used give a broader range of interview length. Although the Study 1 included short question-answer combinations similar to the Connect Four interviews used by Vrij et al. [1], Study 2 included interviews that were significantly longer, many lasting longer than 20 minutes. This gives more detail for the coders to use in making their assessments, which may make them more accurate or instead introduce more variability in judgments that reduces the reliabilities of the coders. Longer interviews also gives the interviewees more time to adapt to the questioning strategy of the interviewers, making their veracity more difficult to detect [12]. Vrij et al. predicted that RJs would be reliable and valid and we also verified the validity in our two studies with varying interview length.

In our replication and extension of Vrij’s work with the RJ method, we also make the following predictions regarding the utility of RJs for differentiating truth and deception and the ability of minimally trained coders to make accurate truth-deception judgments after they have made their RJs:

H1: Deceivers will demonstrate fewer of the following RJ criteria than truth-tellers: Quantity of details, contextual embedding, reproduction of conversation, visual details, auditory details, temporal details, and spontaneous corrections.

One advantage of using third-party coders is that they are able to be more objective and thus, more accurate than the interviewers themselves [6]. They are not as cognitively busy and when they use RJs, they should be more accurate than the interviewers who are making truth-deception judgments.

H2: Coders using RJs make truthful estimates that are more accurate than participant-interviewers.

4. Method

4.1. Participants and Procedure

Eight observers, two males and six females aged 20–23, participated in the study. All were undergraduate communication majors who participated for independent study credit. Each observer coded between 12 and 30 videos from each dataset and each video was coded by two observers. Observers coded independently on laptop computers with headphones using the (C# Behavioral Annotation System (C-BAS)) [42]. C-BAS is a computer program that brings up a window in which coders can see the video they are coding, the coding system, and their codes simultaneously. They were instructed to watch the entire video segment and then rate the interviewee on the 7 RJs listed above on 7-point Likert scales ranging from (1) absent to (5) very much present. They then rated the truthfulness of the participant using a 7-point scale.

4.2. Stimulus Videos

We selected 70 interviews from the Study 1 and 84 interviews from Study 2 described above to comprise the current sample. In the Study 1 [40], participants lied or told the truth on six questions in a counter-balanced order when questioned by professional polygraph examiners (see Appendix A). All the lies were sanctioned and a teleprompter instructed the interviewees when to lie and when to tell the truth. Every three questions, the participant and the expert interviewer rated the participant's truthfulness. Accuracy was determined by comparing the interviewer's rating to the participant's self-reported truthfulness on each question. The coders for the current study rated the participant using RJs after each question. In Study 2 [41], participants played a trivia game with a confederate who encouraged cheating in some of the sessions. Professional polygraph examiners interviewed participants after the game to determine cheaters and non-cheaters. The 13 questions that were asked appear in Appendix B. Lies were both sanctioned and unsanctioned because some participants were instructed to lie to the examiner by a researcher and some were not told to lie and so lied on their own. Some of the interviews were conducted via videoconference and some were conducted face-to-face (FtF). The coders for the present study rated the participants' for RJs at the conclusion of the interview.

4.3. Coding and Training

We followed the training method used by Vrij et al. [1] as closely as possible. The first author met with all eight coders to explain the RJ coding system and showed the

coders examples from the video sets of all the RJs (none of the examples were used in the final sample). Then, copies of the CBCA and RM criteria and their descriptions from Vrij's 2008 book ([35]) were distributed to the coders to take home and read. A week later, all eight coders met with the first author again and conducted some sample ratings using the C-BAS program. The instructor wrote all their ratings on a white board and any discrepancies greater than 1 point between the ratings were discussed, often after watching the segment again. Once the first author and the coders felt confident about their judgments and felt that watching further examples was not necessary, the coders were allowed to conduct rating sessions on their own in the lab over the next six weeks. The first author conducted periodic checks of the coders' reliabilities and held periodic meetings with the raters one-on-one to prevent any slippage in reliabilities over time. Overall, the coders spent approximately 90 minutes in training and approximately 15 hours completing the judgment task.

5. Results

Our primary purpose in this study was to determine whether there were any constraints on the reliability and accuracy that could be achieved using RJ coding. First, we were concerned with whether the length of the interview would have an effect on inter-coder reliabilities because the longer interviews in Study 2 could give the coders more data on which to base their decisions but could also create conflicting data that makes a general rating more difficult. For Study 1, each question was rated by two different coders, so intraclass correlations (ICC) were computed to assess the inter-coder reliability for each question. The ICC is the correlation between one measurement on a target and another measurement obtained on that target and is used to establish a correlation between pairs of observations that do not have an obvious order [43]. Shrout and Fleiss discuss six different versions of the ICC and most can be calculated using the reliability analysis function in SPSS. We used the one-way random version of the ICC with a 95% confidence interval and a test value of 0. The reliabilities, by question, for Study 1 are found in Table 2. The minimum acceptable ICC for inter-coder reliability is .60 [43]. Those RJs that achieved acceptable reliability between two coders are bolded in Table 2. Acceptable reliabilities were achieved for 60% of the ratings, but a quick glance at Table 2 reveals that Question 2 was especially problematic because acceptable reliabilities were not achieved on that question for any of the 7 RJ codes. That question, "What would you do if your boss gave you credit for someone else's work?" was the only one that was hypothetical rather than factual in nature, suggesting that perhaps the type of question affects the reliability of the coding. Also, the auditory details criteria achieved acceptable reliabilities on just one question (Q3), suggesting that it posed the greatest problem for the coders,

perhaps because auditory details came up very rarely in these interviews (the ratings ranged from 1 to 4 on a 7-point scale, $M = 1.19$, $SD = .42$).

For the Study 2, ratings were made after watching the entire interview during most of which, all the participants were telling the truth. Only later in the interview, when participants were asked about the cheating during the trivia game did differences emerge between the cheaters and the non-cheaters. Thus, coders were asked to wait until they had seen the entire interview before making their ratings. The reliabilities for the cheating study are found in Table 3 and were also virtually all in the acceptable range. The exception was auditory details which proved problematic for coders probably because they occurred so rarely in the interviews ($M = 1.35$, $SD = .72$). For comparison, we collapsed Study 1 into two blocks of questions because participants were prompted to answer truthfully or deceptively in blocks of three questions in a counter-balanced order. We dropped Q2 from Block 1 because it proved problematic for coders and was a different question type from the other questions (hypothetical rather than factual) and the reliabilities for Block 1 improved. Table 3 reveals that Study 2 reliabilities were comparable to the Study 1 reliabilities and were mostly in the acceptable range.

Hypothesis one predicted that deceivers would demonstrate fewer of the following RJ criteria than truth-tellers: Quantity of details, contextual embedding, reproduction of conversation, visual details, auditory details, temporal details, and spontaneous corrections. In order to test this hypothesis, we conducted a MANOVA for the Study 1 data with question block (question 2 removed) and truth-deception order (TD Order) as the independent variables and the RJs and truthfulness ratings as the dependent variables. Because the participants were truthful or deceptive in blocks of three questions, we would expect a significant block X TD order interaction effect. Deceivers showing fewer of the RJ criteria would confirm the hypothesis.

We found a significant main effect for block for quantity of details $F(1,71) = 20.91$, $p < .001$, $\eta^2 = .29$ and contextual embedding $F(1,71) = 5.16$, $p = .03$, $\eta^2 = .003$, both of which showed greater detail and contextual embedding in block 2 compared to block 1 regardless of whether they were truth-tellers or deceivers. We also found a significant block X TD order interaction effect for visual details $F(1,71) = 11.28$, $p = .001$, $\eta^2 = .08$ displayed in Figure 1. Contrary to the hypothesis, the results show more visual details being used by deceivers than truth-tellers in block 1 and no difference in block 2. The results also revealed a significant block X TD order interaction effect for temporal details $F(1,71) = 3.54$, $p = .05$, $\eta^2 = .05$. Consistent with the hypothesis, the means displayed in Figure 2 show that deceivers used fewer temporal details than truth-tellers in the second block. For the overall ratings of truthfulness, the analysis revealed a significant block X

TD order interaction effect as well $F(1,71) = 3.51$, $p = .05$, $\eta^2 = .05$. There was no difference in truthfulness ratings between truth-tellers and deceivers in block 1 but in block 2, truth-tellers were rated significantly more truthful than deceivers, indicating that coders were better able to rate truthfulness in block 2 than block 1 (see Figure 3). Those who began with truth in block 1 became much more transparent when they switched to deception in block 2 than those who began with deception and then switched to truth in block 2.

For Study 2, we conducted a MANOVA with modality (videoconference vs. face-to-face) and truth-deception (TD) as the independent variables and the 7 RJs and truthfulness ratings as the dependent variables (we collapsed the sanctioned and unsanctioned deceivers into a single condition due to the small sample size). The results revealed modality main effects for reproduction of conversation $F(1, 80) = 5.92$, $p = .02$, $\eta^2 = .07$ (FTF: $M = 2.05$, $SD = .87$; CMC: $M = 2.80$, $SD = 1.63$), auditory details $F(1, 80) = 5.94$, $p = .02$, $\eta^2 = .07$ (FTF: $M = 1.19$, $SD = .41$; CMC: $M = 1.48$, $SD = .60$), and a near-significant main effect for temporal details $F(1, 80) = 3.11$, $p = .08$, $\eta^2 = .04$ (FTF: $M = 2.08$, $SD = .80$; CMC: $M = 2.46$, $SD = 1.42$). Participants conversing in CMC used more of these RJ criteria than those conversing via FtF.

The results also revealed a modality X TD interaction effect $F(1, 80) = 3.59$, $p = .05$, $\eta^2 = .04$ for visual details (see Figure 4). Truth-tellers used more visual details than deceivers when conversing FtF but deceivers and truth-tellers were virtually indistinguishable in CMC. A near-significant interaction effect for spontaneous corrections $F(1, 80) = 3.22$, $p = .08$, $\eta^2 = .04$ indicated that truth-tellers made more corrections in FtF but were similar to deceivers in CMC. Both of these results are consistent with H1.

Table 2: ICC reliabilities for the 6 questions in Study 1.

Code	Q1 What's the worst job you've ever had and why?	Q2 What would you do if your boss gave you credit for someone else's work?	Q3 Tell me about a time when you thought about stealing something valuable	Q4 Please tell me everything you did today from when you woke up until when you arrived here at the lab.	Q5 What do you consider your greatest strengths?	Q6 What else will you do today?
Quantity of Details	.72	.55	.80	.77	.80	.78
Contextual Embedding	.04	-.32	.75	.80	.86	.77
Reproduction of Conversation	.87	.43	.74	.54	.60	.50
Visual Details	.70	-.10	.82	.65	.54	.67
Auditory Details	.56	-.14	.67	.36	-.13	.40
Temporal Details	.30	.02	.60	.68	.72	.55
Spontaneous Corrections	.65	.46	.69	.74	.73	.74

Table 3: ICC reliabilities for Study 1 and Study 2 in question blocks.

Code	Study 2	Study 1 Block 1	Study 1 Block 1 w/o Q2	Study 1 Block 2
Quantity of Details	.74	.72	.72	.80
Contextual Embedding	.75	.39	.33	.84
Reproduction of Conversation	.80	.82	.86	.61
Visual Details	.78	.73	.78	.63
Auditory Details	.31	.56	.70	.21
Temporal Details	.72	.40	.51	.60
Spontaneous Corrections	.77	.65	.70	.80

Figure 1: Significant Block X TD Order Interaction Effect for Study 1 visual details

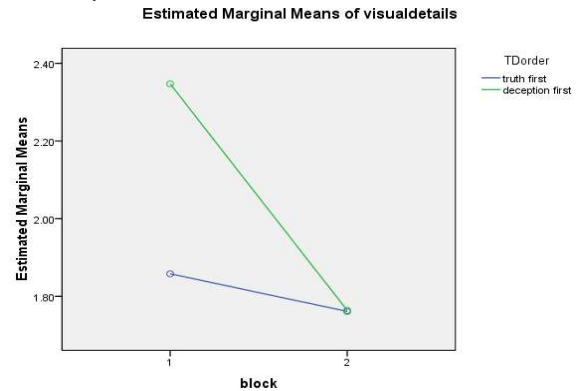


Figure 2: Significant Block X TD Order Interaction Effect for Study 1 temporal details

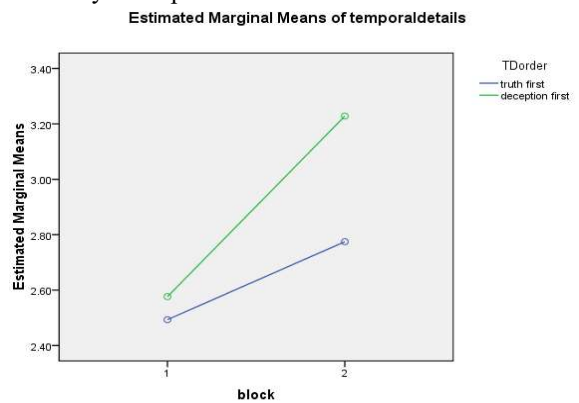


Figure 3: Significant Block X TD Order Interaction Effect for Study 1 truth ratings

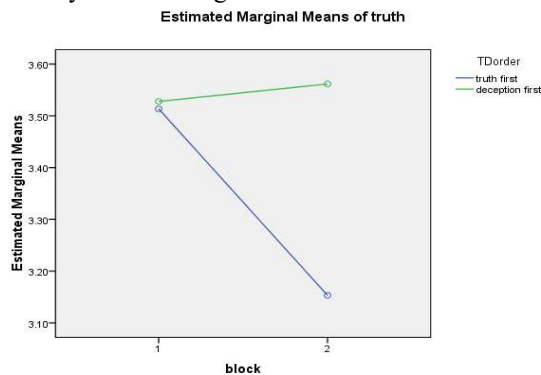
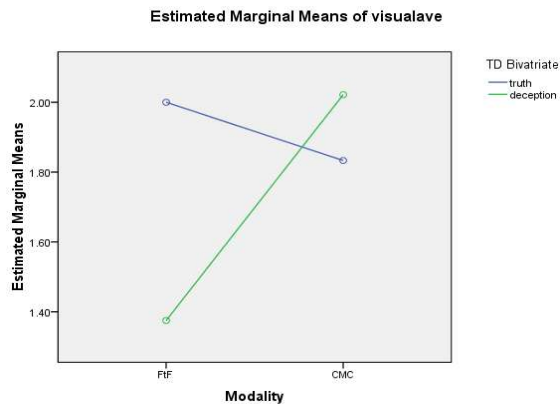


Figure 4: Modality X TD interaction effect for Study 2 visual details.



Hypothesis two predicted that coders using RJs should make truthful estimates that are more accurate than the participant-interviewers. In the case of the Study 1, Burgoon et al. reported total accuracy of their expert interviewers at 60.5% with a truth accuracy of 73% and a lie accuracy of 48%. Because the interviewers and participants rated truthfulness on a scale, we asked the RJ coders to also rate the truthfulness of the participants on a scale rather than make dichotomous truth-deception judgments. To determine the accuracy rates, we then converted those ratings into truth-deception judgments by designating any value below the midpoint of the truthfulness scale as deceptive and any value above the midpoint of the scale as truthful. If an answer was rated at the midpoint of the scale, it was considered neither truthful nor deceptive and was treated as missing data in the analyses. This was the case in a small number of interviews (between 4-15 coded responses depending on the question). As Table 4 reveals, the accuracy of the RJ coders was lower than the experts who made their ratings during the interviews.

For Study 2, Dunbar et al. [41] reported 58% accuracy overall for their expert interviewers with a truth accuracy of 60% and a lie accuracy (with the sanctioned and unsanctioned conditions combined) of 54%. The results in Table 4 demonstrate that the coders of the cheating study had an accuracy rate of 67.8% which is an improvement over the expert interviewers but is largely due to their improved truth accuracy.

Table 4: Accuracy Rates for the CB Study

	Study 1 Questions							
	Q1	Q2	Q3	Q4	Q5	Q6	Ave Study 1	Study 2
Number of Correct Judgments	32	29	25	31	37	31	30.8	53
Valid N	60	59	62	62	60	58	60.2	78
Percent Correct	53.3%	49.1%	40%	50%	61.7%	53.4%	51.3%	67.9%
Truth Accuracy	62.2%	35.1%	43.3%	63.9%	63.9%	55.6%	54.0%	75.6%
Lie Accuracy	25%	44.4%	25%	21.6%	37.8%	29.7%	30.6%	51.2%

6. Discussion

Rapid judgments are an important innovation in the improvement of deception detection and have several advantages over the traditional use of laborious microscopic coding or specialized training for coders. However, before we can trust RJs over other methodologies, their viability must be demonstrated first in the ability of minimally trained coders to reliably judge them and then in the ability of RJs to accurately discriminate between truth and deception and lead observers to make accurate deceptiveness estimates.

One advantage of RJs is that they can be used in deception detection with no need for written transcripts or specialized equipment, and minimal training time for coders but this is only an advantage when the coders can make reliable judgments. The coders in this study spent less than two hours in training and still achieved acceptable reliabilities even over long segments of interaction. The ICCs presented in Table 3 comparing the two studies demonstrate that the reliabilities of RJs in Study 1 which had short question-and-answer segments were comparable to the longer interview-length judgments in Study 2. Generally, the ICCs revealed that RJs can be made reliably with minimal training but that the type of questions (i.e. hypothetical vs. factual) and the type of detail (i.e. auditory vs. spacial) may make a difference and should be investigated further.

Hypothesis one stated that deceivers would demonstrate fewer of the following RJ criteria than truth tellers: quantity of details, contextual embedding, reproduction of conversation, visual details, auditory cues, temporal details, and spontaneous corrections. Hypothesis one was not supported because deceivers used more of some types of details (visual) but fewer of other types of details (temporal) and still other RJs were not significantly different between truth and deception. Vrij [35] cites some inconsistent findings for the diagnosticity of CBCA and RM criteria in other studies suggesting that more research into the nuances of when RJs are the most diagnostic is still needed. Some of the variations in the directionality of RJs may have been a result of the differences in the questions asked in our two studies. Previous studies of RJs were focused on questions about a particular event but our questions ranged from personal history questions to accusatory questions. The questions in block 1 of Study 1, for example, were very personal whereas block 2 questions asked about mundane daily activities. In fact, those individuals, regardless of whether they were deceivers or truth-tellers provided more details and contextual embedding than block 1. This is not only because block 2 questions asked about what they did, but block 1 could have been incriminating to the individual, so the less detail the participant gave, the less detectable deceptions would have been. It is interesting that there were more visual details for deceivers than truth-tellers in block 1 (contrary

to H1). While it is common that deceivers will shut down and not be overly animated, displaying fewer kinesic movements [9, 35], because these questions were negative, it is possible that deceivers were trying to overcompensate for their lying by appearing as if they were telling the truth (being more animated while speaking).

The second hypothesis stated that coders using RJs should make truthful estimates that are more accurate than the participant-interviewers. The interviewers had overall accuracy rates of 60.5% and 58% in Study 1 and Study 2 respectively. For comparison, the accuracy rates of the RJ coders in Study 1 ranged from 40% in Question 1 (“Tell me about a time when you thought about stealing something valuable”) to 61.7% on Question 5 (“What do you consider your greatest strengths?”). The lie accuracy in Study 1 proved especially problematic; the lie accuracy of the interviewers was a less-than-impressive 48% but the lie accuracy of the RJ coders was only 30.6%. The RJ coders fared better in Study 2 with the longer interviews, achieving an overall accuracy rate of 67.9% (10% higher than the interviewers). Therefore, the longer interviews seemed to have increased accuracy rather than hindered it. Longer interviews may have enabled both the participant-interviewer and the coder to have more information and more RJs on which to base their global assessment. In Study 2, the coders not only attained improved detection accuracy over Study 1 but also better than the participant-interviewers in Study 2. This can be attributed to the ability of coders to replay the clip as many times as possible, whereas participant-interviewers cannot rewind a live interaction. This could have increased the accuracy between coders and participant-interviewers in Study 2 although it is not known how many of the coders really used the rewinding and replaying features that were available to them. Future research should investigate whether observers watching live interactions (and thus making judgments under the same conditions as the participant-interviewers) are more accurate than the participant-interviewers themselves.

7. Conclusion

The premise of the RJs approach to coding is that when people are asked to pay attention to verbal and nonverbal cues rather than attempt to detect deceit, they are more accurate in their judgments [38, 39]. Initial testing of RJs as a method of deception detection proved promising. In this study, two datasets were used to validate the reliability and accuracy of RJs as a viable method of detection deception. We found that minimally trained coders were indeed able to reliably make RJs after watching both long and short segments of interactions. The coders who viewed longer segments were more accurate than both the coders who viewed shorter segments and the participant-interviewers. The RJs did not discriminate between truth and deception in the manner that was hypothesized which

raises more questions about the conditions under which making RJs from verbal and nonverbal cues achieves accurate detection of veracity. Future research could profitably isolate what conditions are most amenable to the RJ approach and what cues are likely to be the most diagnostic for judges with limited training.

8. References

1. A. Vrij, et al., "Rapid judgements in assessing verbal and nonverbal cues: Their potential for deception researchers and lie detection," *Applied Cognitive Psychology*, vol. 18, no. 3, 2004, pp. 283-296.
2. C.F. Bond and B.M. DePaulo, "Accuracy of Deception Judgments," *Personality and Social Psychology Review*, vol. 10, no. 3, 2006, pp. 214-234.
3. T. Levine, "A few transparent liars," *Communication Yearbook*, vol. 34, 2010, pp. 41-61.
4. P. Ekman, et al., "A few can catch a liar," *Psychological Science*, vol. 10, no. 3, 1999, pp. 263-266.
5. A. Vrij, "Why professionals fail to catch liars and how they can improve," *Legal and Criminological Psychology*, vol. 9, 2004, pp. 159-181; DOI 10.1348/1355325041719356.
6. N.E. Dunbar, et al., "The effects of participation on the ability to judge deceit," *Communication Reports*, vol. 16, no. 1, 2003, pp. 23-34.
7. G. Giordano, et al., "Reviewers and the Detection of Deceptive Information in Recorded Interviews," *Journal of Applied Social Psychology*, vol. 41, no. 2, 2011, pp. 252-269.
8. M.L. Jensen, et al., "Effects of automated and participative decision support in computer-aided credibility assessment," *Journal of Management Information Systems*, vol. 28, no. 1, 2011, pp. 203-236.
9. B. DePaulo, et al., "Cues to deception," *Psychological Bulletin*, vol. 129, no. 1, 2003, pp. 74-118.
10. M. Zuckerman and R. Driver, "Telling lies: Verbal and nonverbal correlates of deception," *Nonverbal Communication: An Integrated Perspective*, A. W. Siegman and S. Feldstein, eds., Erlbaum, 1985, pp. 129-147.
11. M.L. Jensen, et al., "Automatic, multimodal evaluation of human interaction," *Group Decision and Negotiation*, vol. 19, no. 4, 2010, pp. 367-389.
12. J.K. Burgoon and T. Qin, "The dynamic nature of deceptive verbal communication," *Journal of Language and Social Psychology*, vol. 25, no. 1, 2006, pp. 76.
13. A. Vrij, et al., "The effect of informing liars about Criteria-Based Content Analysis on their ability to deceive CBCA-raters," *Legal and Criminological Psychology*, vol. 5, no. 1, 2000, pp. 57-70.
14. L. Anolli and R. Ciceri, "The voice of deception: Vocal strategies of naive and able liars," *Journal of Nonverbal Behavior*, vol. 21, no. 4, 1997, pp. 259-284.
15. A. Elkins and J. Stone, "The effect of cognitive dissonance on argument language and vocalics," *Book The effect of cognitive dissonance on argument language and vocalics*, Series The effect of cognitive dissonance on argument language and vocalics, ed., Editor ed.^eds., 2011, pp.
16. P. Rockwell, et al., "Measurement of deceptive voices: Comparing acoustic and perceptual data," *Applied Psycholinguistics*, vol. 18, no. 4, 1997, pp. 471-484; DOI 10.1017/s01427164000/0948.
17. A. Vrij and S. Heaven, "Vocal and verbal indicators of deception as a function of lie complexity," *Psychology, Crime and Law*, vol. 5, no. 3, 1999, pp. 203-215.
18. D.C. Derrick, et al., "Border Security Credibility Assessments via Heterogeneous Sensor Fusion," *IEEE Intelligent Systems*, vol. 25, no. 3, 2010, pp. 41-49.
19. D.A. Pollina, et al., "Comparison of Polygraph Data Obtained From Individuals Involved in Mock Crimes and Actual Criminal Investigations," *Journal of Applied Psychology*, vol. 89, no. 6, 2004, pp. 1099-1105.
20. J.T. Hancock, et al., "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication," *Discourse Processes*, vol. 45, no. 1, 2008, pp. 1-23.
21. A. Vrij, et al., "Outsmarting the liars: Toward a cognitive lie detection approach," *Current Directions in Psychological Science*, vol. 20, no. 1, 2011, pp. 28.
22. R. Bull, "Can training enhance the detection of deception," *Credibility Assessment*, J. C. Yuille, ed., Kluwer Academic Publishers, 1989.
23. C.J. Dando and R. Bull, "Maximising opportunities to detect deception: Training police officers to interview tactically," *Journal of Investigative Psychology and Offender Profiling*, vol. 8, 2011, pp. 189-202; DOI 10.1002.jip.145.
24. P. Ekman, et al., "Body movement and voice pitch in deceptive interaction," *Semiotica*, vol. 16, 1976, pp. 23-27; DOI 10.1515/semi.1976.16.1.23.
25. M.G. Frank and T.H. Feeley, "To catch a liar: Challenges for research in lie detection training," *Journal of Applied Communication Research*, vol. 31, no. 1, 2003, pp. 58-75.
26. J. Masip, et al., "Training to detect what? The biasing effects of training on veracity judgments," *Applied Cognitive Psychology*, vol. 23, 2009, pp. 1282-1296; DOI 10.1002/acp.1535.
27. T.R. Levine, et al., "Testing the effects of nonverbal behavior training on accuracy in deception detection with the inclusion of a bogus training control group," *Western Journal of Communication*, vol. 69, no. 3, 2005, pp. 203-217.
28. K. Fiedler, "Training lie detectors to use nonverbal cues instead of global heuristics," *Human Communication Research*, vol. 20, no. 2, 1993, pp. 199-223.
29. M.A. deTurck, "Training observers to detect spontaneous deception: effects of gender," *Communication Reports*, vol. 4, 1991, pp. 81-89; DOI doi: 10.1080/08934219109367528.
30. M.A. deTurck, et al., "The effects of training social perceivers to detect deception from behavioral cues," *Communication Quarterly*, vol. 38, 1990, pp. 189-199; DOI 10.1080/014633790093697353.
31. B.M. DePaulo and W.L. Morris, "Discerning lies from truths: Behavioral cues to deception and the indirect pathway of intuition," *The detection of deception in forensic contexts*, P. A. Granhag and L. A. Stromwall, eds., Cambridge University Press, 2004, pp. 15-40.
32. A. Vrij, "Implicit lie detection," *The Psychologist*, vol. 14, 2001, pp. 58-60.
33. A. Vrij, et al., "Police officer's ability to detect deceit: The benefits of indirect deception detection," *Legal and Criminological Psychology*, vol. 6, no. 2, 2001, pp. 185-196.

34. M. O'Sullivan, et al., "The effect of comparisons on detecting deceit," *Journal of Nonverbal Behavior*, vol. 12, no. 3, 1988, pp. 203-215.
35. A. Vrij, *Detecting lies and deceit: Pitfalls and opportunities*, Wiley, 2008.
36. The Global Deception Research Team, "A world of lies," *Journal of Cross-Cultural Psychology*, vol. 37, no. 1, 2006, pp. 60-74.
37. L. Anolli, et al., "Linguistic styles in deceptive communication: Dubitative ambiguity and elliptic eluding in packaged lies," *Social Behavior and Personality*, vol. 31, no. 7, 2003, pp. 687-710.
38. S. Porter, et al., "Truth, lies, and videotape: An investigation of the ability of federal parole officers to detect deception," *Law and Human Behavior*, vol. 24, no. 6, 2000, pp. 643-658.
39. A. Vrij, *Detecting Lies and Deceit: The Psychology of Lying and the Implications for Professional Practice*, John Wiley & Sons Ltd, 2000.
40. J.K. Burgoon, et al., *Rapid credibility assessment: A report to the counterintelligence field activity*, Counterintelligence Field Activity (Contract No. H9C 104-07-C-0011), 2008.
41. N. Dunbar, et al., "Cheating and credibility: How modality and power affect veracity detection," *Book Cheating and credibility: How modality and power affect veracity detection*, Series Cheating and credibility: How modality and power affect veracity detection, ed., Editor ed.^eds., 2010, pp.
42. T.O. Meservy, *CBAS 2.0: Software Tools to Code Human Nonverbal Behavior*, Center for Identification Technology Research Spring Conference, 2010.
43. J.L. Fleiss, "Balanced incomplete block designs for inter-rater reliability studies," *Appl. Psychol. Meas.*, vol. 5, no. 1, 1981, pp. 105-112.